STAT 538 Final Presentation

Surprises in High-Dimensional Ridgeless Least Squares Interpolation

Paper Authors:
Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani

Paper link:
https://doi.org/10.1214/21-AOS2133 (The Annals of Statistics, 2022)

March 6, 2023

Presented by *Yikun Zhang*

A Central Tenet in Machine Learning

Assume that $y_i = f(x_i) + \epsilon_i$ with $(x_i, \epsilon_i) \sim P_x \times P_\epsilon$ for $i = 1, ..., n$.

- Training Risk: $\frac{1}{n}\sum_{i=1}^{n} L\left(h(x_i), y_i\right)$ for some loss function L.
- Test Risk: $\mathbb{E}_{(x,y)\sim P_{xy}}\left[L(h(x), y)\right]$.

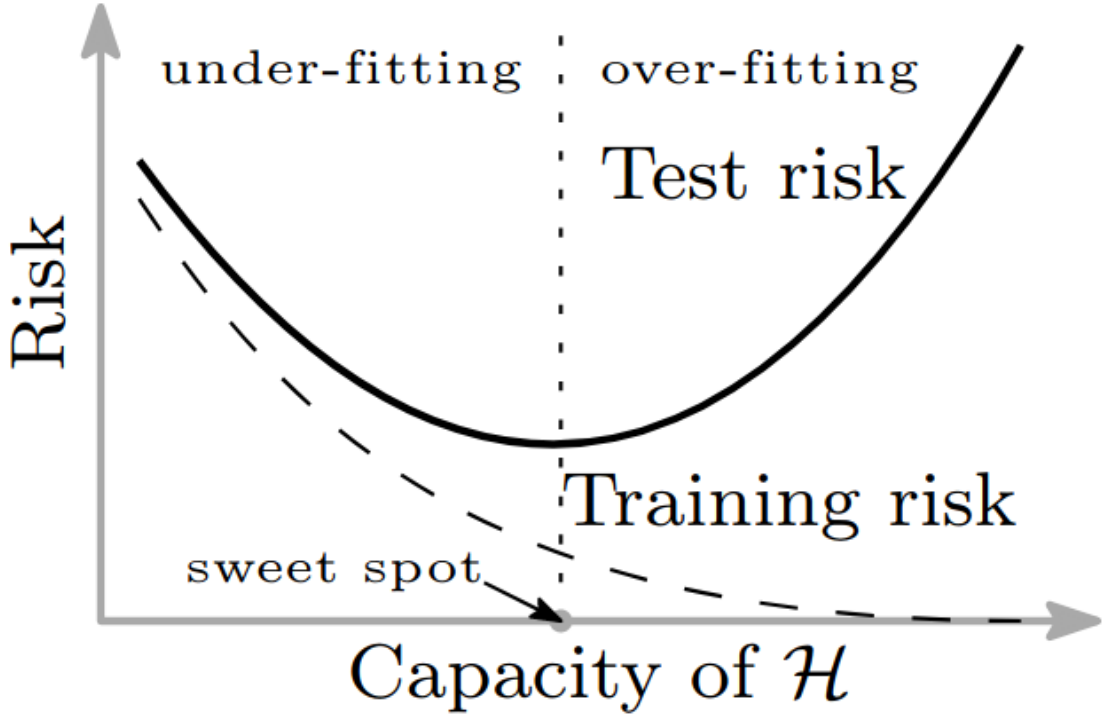
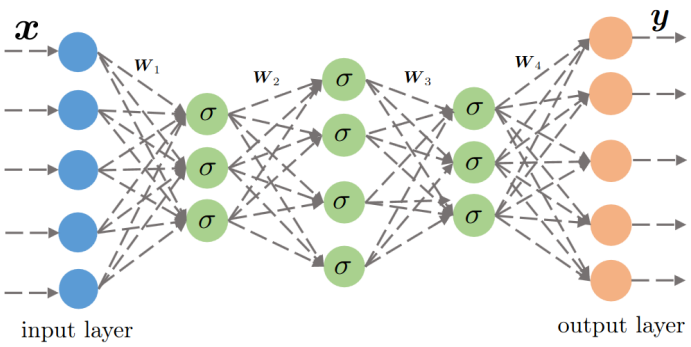


Figure 1: Classical bias-Variance trade-off (Belkin et al., 2019).

Contradictory Evidence in Deep Neural Networks



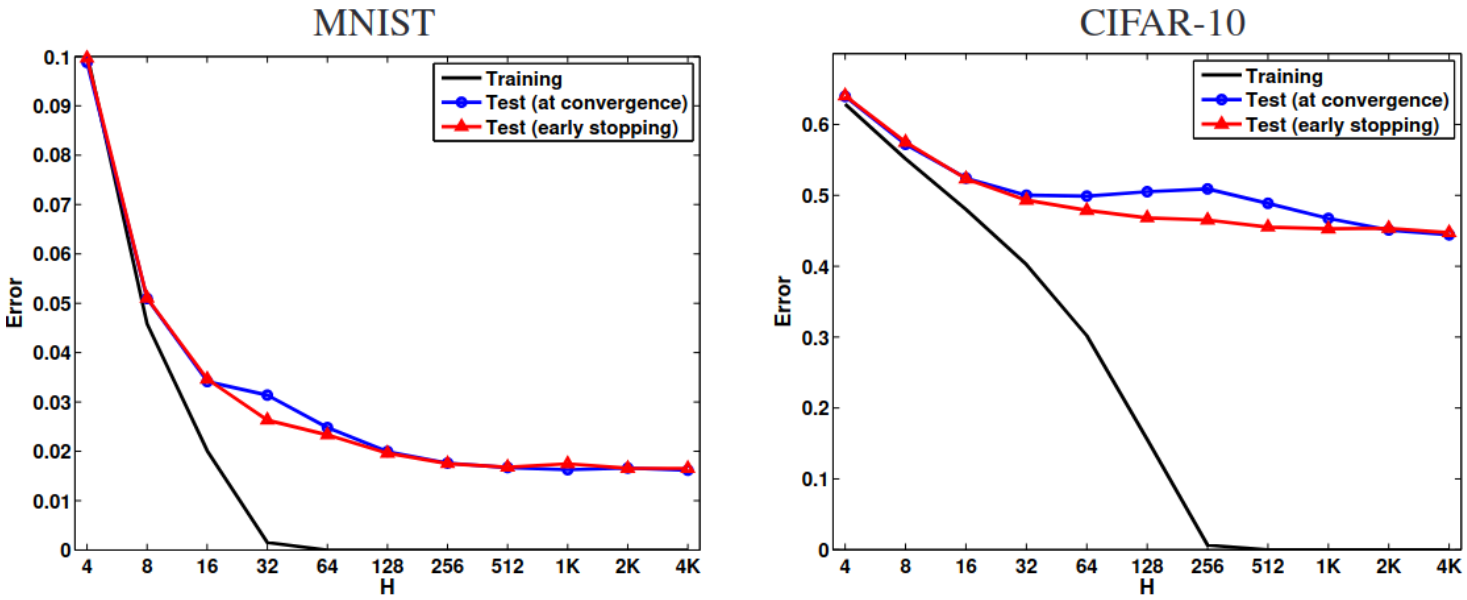


Figure 2: Training and test errors of two-layer Neural Networks (NNs) with different number of hidden units H (Neyshabur et al., 2014).

- Notes: The number of parameters is $H(d+K)$ for each two-layer NNs, where d is the number of features and K is the size of the output layer.

Interpolating/Overparameterized Regime

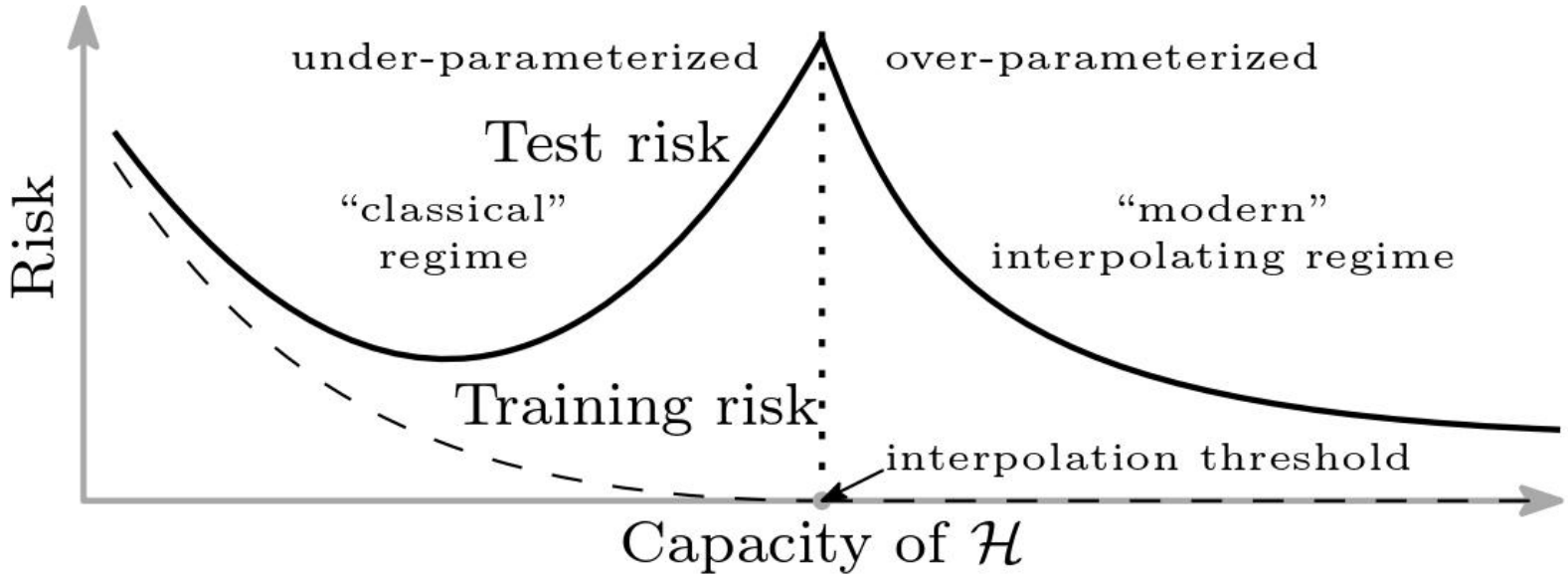


Figure 3: An extension of the classical bias-variance trade-off framework: the "double descent" risk curve (Belkin et al., 2019).

Overparametrized Linear Models

Data: $\{(y_i, x_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ from the linear model

$$y_i = x_i^T \beta + \epsilon_i \quad \text{with} \quad (x_i, \epsilon_i) \overset{\text{i.i.d.}}{\sim} \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x_i) = 0, \mathrm{Cov}(x_i) = \Sigma$, and $\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$.

Overparametrized Linear Models

Data: $\{(y_i, x_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ from the linear model

$$y_i = x_i^T \beta + \epsilon_i \quad \text{with} \quad (x_i, \epsilon_i) \overset{\text{i.i.d.}}{\sim} \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x_i) = 0, \mathrm{Cov}(x_i) = \Sigma$, and $\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$.

- *Isotropic features:* $\Sigma = \boldsymbol{I}_p$.
- *Latent space features:* $\Sigma = WW^T + \boldsymbol{I}_p$ with $W \in \mathbb{R}^{p \times d}, d \ll p$ and β lies in the span of the columns of W.
- *Nonlinear features:* $x_i = \varphi(Wz_i)$ with $W \in \mathbb{R}^{p \times d}$, $z_i \sim N(0, \boldsymbol{I}_d)$, and φ is a nonlinear activation function.

Overparametrized Linear Models

Data: $\{(y_i, x_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ from the linear model

$$y_i = x_i^T \beta + \epsilon_i \quad \text{with} \quad (x_i, \epsilon_i) \overset{\text{i.i.d.}}{\sim} \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x_i) = 0, \mathrm{Cov}(x_i) = \Sigma$, and $\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$.

- *Isotropic features:* $\Sigma = \boldsymbol{I}_p$.
- *Latent space features:* $\Sigma = WW^T + \boldsymbol{I}_p$ with $W \in \mathbb{R}^{p \times d}, d \ll p$ and β lies in the span of the columns of W.
- *Nonlinear features:* $x_i = \varphi(Wz_i)$ with $W \in \mathbb{R}^{p \times d}$, $z_i \sim N(0, \boldsymbol{I}_d)$, and φ is a nonlinear activation function.

Question: Why do we study overparametrization on simple linear models?

Connecting Linear Models to Neural Networks

"Lazy training" regime (Geiger et al., 2020): model parameter $\theta = (a_i, W_i; i = 1, ..., N)$ stays close to the initialization θ_0 as $\theta = \theta_0 + \Delta$, and we approximate the two-layer neural network model

$$f(z;\theta) \equiv f(z;\boldsymbol{a},\boldsymbol{W}) = \sum_{i=1}^{N} a_i \cdot \varphi\left(w_i^T z\right) \qquad \text{with } a_i \in \mathbb{R}, w_i \in \mathbb{R}^p$$

by

$$\begin{aligned} f(z;\theta) &\approx f(z;\theta_0) + \nabla f(z;\theta_0)^T \Delta \\ &\approx f(z;\theta_0) + \nabla_{\boldsymbol{a}} f(z;\boldsymbol{a}_0,\boldsymbol{W}_0)^T \Delta \boldsymbol{a} + \nabla_{\boldsymbol{W}} f(z;\boldsymbol{a}_0,\boldsymbol{W}_0)^T \Delta \boldsymbol{W} \\ &= f(z;\theta_0) + \underbrace{\sum_{i=1}^{N} \Delta a_i \cdot \varphi\left(w_{0,i}^T x_i\right)}_{\text{Random feature model}} + \underbrace{\sum_{i=1}^{N} a_{0,i} z^T \Delta w_i \cdot \varphi'\left(w_{0,i}^T z\right)}_{\text{Neural tangent kernel model}}. \end{aligned}$$

- The approximation is still nonlinear in the input z but linear in the parameter $\beta = \Delta$.
- The above arguments can be made rigorous in Jacot et al. (2018); Du et al. (2019); Allen-Zhu et al. (2019).

Minimum ℓ_2-Norm Least Squares Regression

Overparametrization ratio: $\gamma := \frac{p}{n} \in (0, \infty)$.

Given the training data $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ and $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$, we solve for the usual least squares estimator $\widehat{\beta} = (X^T X)^{-1} X^T Y$ when $\gamma \leq 1$ (rigorously, X need to have full column rank).

Minimum ℓ_2-Norm Least Squares Regression

Overparametrization ratio: $\gamma := \frac{p}{n} \in (0, \infty)$.

Given the training data $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ and $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$, we solve for the usual least squares estimator $\widehat{\beta} = (X^T X)^{-1} X^T Y$ when $\gamma \leq 1$ (rigorously, X need to have full column rank).

Question: What if $\gamma > 1$ with an underdetermined system of linear equations $Y = X\beta$?

Minimum ℓ_2-Norm Least Squares Regression

Overparametrization ratio: $\gamma := \frac{p}{n} \in (0, \infty)$.

Given the training data $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ and $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$, we solve for the usual least squares estimator $\widehat{\beta} = (X^T X)^{-1} X^T Y$ when $\gamma \leq 1$ (rigorously, X need to have full column rank).

Question: What if $\gamma > 1$ with an underdetermined system of linear equations $Y = X\beta$?

Minimum ℓ_2-norm least squares regression:

$$\widehat{\beta} = \arg\min \left\{ ||b||_2 : b \text{ minimizes } ||Y - Xb||_2^2 \right\}.$$

Solving Minimum ℓ_2-Norm Least Squares Regression

When $\gamma > 1$, we can solve the minimum ℓ_2-norm least squares regression $\widehat{\beta} = \arg\min \left\{ ||b||_2 : b \text{ minimizes } ||Y - Xb||_2^2 \right\}$ by two different methods:

Solving Minimum ℓ_2-Norm Least Squares Regression

When $\gamma > 1$, we can solve the minimum ℓ_2-norm least squares regression $\widehat{\beta} = \arg\min \left\{ ||b||_2 : b \text{ minimizes } ||Y - Xb||_2^2 \right\}$ by two different methods:

1. **Gradient descent:** $\beta^{(k+1)} \leftarrow \beta^{(k)} + tX^T \left(Y - X\beta^{(k)} \right), k = 0, 1, ...,$ where $\beta^{(0)} = 0$ and $t \in \left(0, \frac{1}{\lambda_{\max}(X^T X)} \right)$ with $\lambda_{\max}(X^T X)$ being the largest eigenvalue of $X^T X$.

Solving Minimum ℓ_2-Norm Least Squares Regression

When $\gamma > 1$, we can solve the minimum ℓ_2-norm least squares regression $\widehat{\beta} = \arg\min \left\{ ||b||_2 : b \text{ minimizes } ||Y - Xb||_2^2 \right\}$ by two different methods:

1. **Gradient descent:** $\beta^{(k+1)} \leftarrow \beta^{(k)} + tX^T \left(Y - X\beta^{(k)}\right), k = 0, 1, ...,$ where $\beta^{(0)} = 0$ and $t \in \left(0, \frac{1}{\lambda_{\max}(X^T X)}\right)$ with $\lambda_{\max}(X^T X)$ being the largest eigenvalue of $X^T X$.

2. **Analytic solution ("Ridgeless"):** Consider the ridge regression

$$\widehat{\beta}_\lambda = \arg\min_{b \in \mathbb{R}^p} \left[\frac{1}{n} ||Y - Xb||_2^2 + \lambda ||b||_2^2 \right] = \left(X^T X + n\lambda \boldsymbol{I}\right)^{-1} X^T Y \overset{(*)}{=} X^T \left(XX^T + n\lambda \boldsymbol{I}\right)^{-1} Y, \tag{1}$$

and $\widehat{\beta} = \lim_{\lambda \to 0^+} \widehat{\beta}_\lambda$, where we use the "kernel tricks" in (*).

Out-of-Sample Prediction Risk

We evaluate the minimum ℓ_2-norm least squares regression through the out-of-sample prediction risk with $x_0 \sim P_X$ as:

$$
\begin{aligned}
R_X\left(\widehat{\beta};\beta\right) &= \mathbb{E}\left[\left(x_0^T\widehat{\beta} - x_0^T\beta\right)|X\right] \\
&= \mathbb{E}\left[\left|\left|\widehat{\beta}-\beta\right|\right|_\Sigma^2 |X\right] \\
&= \underbrace{\left|\left|\mathbb{E}(\widehat{\beta}|X)-\beta\right|\right|_\Sigma^2}_{B_X(\widehat{\beta};\beta)} + \underbrace{\text{Trace}\left[\text{Cov}(\widehat{\beta}|X)\Sigma\right]}_{V_X(\widehat{\beta};\beta)},
\end{aligned}
$$

where $||x||_\Sigma^2 = x^T\Sigma x$.

Out-of-Sample Prediction Risk

We evaluate the minimum ℓ_2-norm least squares regression through the out-of-sample prediction risk with $x_0 \sim P_X$ as:

$$\begin{aligned} R_X\left(\widehat{\beta};\beta\right) &= \mathbb{E}\left[\left(x_0^T\widehat{\beta} - x_0^T\beta\right)|X\right] \\ &= \mathbb{E}\left[\left|\left|\widehat{\beta}-\beta\right|\right|_\Sigma^2 |X\right] \\ &= \underbrace{\left|\left|\mathbb{E}(\widehat{\beta}|X)-\beta\right|\right|_\Sigma^2}_{B_X(\widehat{\beta};\beta)} + \underbrace{\text{Trace}\left[\text{Cov}(\widehat{\beta}|X)\Sigma\right]}_{V_X(\widehat{\beta};\beta)}, \end{aligned}$$

where $||x||_\Sigma^2 = x^T\Sigma x$.

If we write the minimum ℓ_2-norm least squares estimator as $\widehat{\beta} = (X^TX)^+X^TY$ with $(X^TX)^+$ being the pseudoinverse of X^TX, then

$$B_X(\widehat{\beta};\beta) = \beta^T\Pi\Sigma\Pi\beta \quad \text{and} \quad V_X(\widehat{\beta};\beta) = \frac{\sigma^2}{n}\text{Trace}(\widehat{\Sigma}^+\Sigma),$$

where $\widehat{\Sigma} = \frac{X^TX}{n}$ and $\Pi = \boldsymbol{I} - \widehat{\Sigma}^+\widehat{\Sigma}$; see Lemma 1 in Hastie et al. (2022).

Underparametrized Asymptotics ($\gamma < 1$)

Under the linear model setting,

$$y = x^T\beta + \epsilon \quad \text{with} \quad (x, \epsilon) \sim \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x) = 0, \mathrm{Cov}(x) = \Sigma$, and $\mathbb{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2$.

If x has a finite 4-th moment and $\lambda_{\min}(\Sigma) \geq c > 0$ for some constant c, then as $n, p \to \infty$ with $p/n \to \gamma < 1$,

$$\lim_{n\to\infty} R_{\mathrm{X}}(\widehat{\beta}, \beta) = \sigma^2 \frac{\gamma}{1-\gamma};$$

see Proposition 2 in Hastie et al. (2022), where the proof leverages the Marchenko-Pastur theorem (Marčenko and Pastur, 1967).

• Notes: In the underparameterized case ($\gamma < 1$),
$B_{\mathrm{X}}(\widehat{\beta}, \beta) = \beta^T \Pi \Sigma \Pi \beta = 0$ because $\Pi = \boldsymbol{I} - \widehat{\Sigma}^{-1}\widehat{\Sigma} = 0$.

Overparametrized Asymptotics (Isotropic Features)

Recall the linear model setting:

$$y = x^T\beta + \epsilon \quad \text{with} \quad (x, \epsilon) \sim \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x) = 0, \mathrm{Cov}(x) = \Sigma$, and $\mathbb{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2$.

Theorem (Theorem 1 in Hastie et al. 2022)

Assume the above linear model, where $x \sim P_x$ has a finite moment of order $4 + \eta$ for some $\eta > 0$ and $\Sigma = \boldsymbol{I}_p$. Let $r^2 = ||\beta||_2^2$. Then, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, it holds (a.s.) that

$$R_X(\widehat{\beta}, \beta) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

Overparametrized Asymptotics (Isotropic Features)

Recall the linear model setting:

$$y = x^T\beta + \epsilon \quad \text{with} \quad (x, \epsilon) \sim \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $\mathbb{E}(x) = 0, \mathrm{Cov}(x) = \Sigma$, and $\mathbb{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2$.

Theorem (Theorem 1 in Hastie et al. 2022)

Assume the above linear model, where $x \sim P_x$ has a finite moment of order $4 + \eta$ for some $\eta > 0$ and $\Sigma = \boldsymbol{I}_p$. Let $r^2 = ||\beta||_2^2$. Then, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, it holds (a.s.) that

$$R_X(\widehat{\beta}, \beta) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

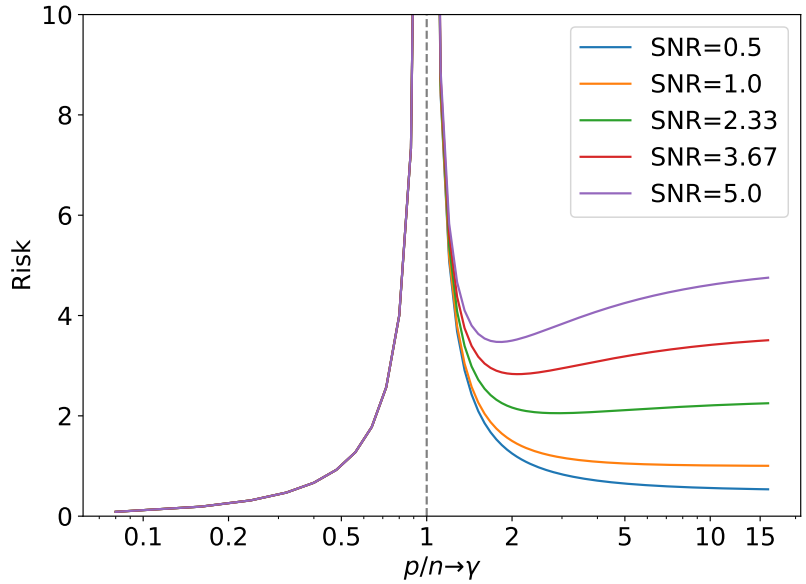
Let $SNR = \frac{r^2}{\sigma^2}$ and note that the risk of the null estimator $\widetilde{\beta} = 0$ is r^2.

- When $\gamma < 1$, $R_X(\widehat{\beta}, \beta) < R_X(\widetilde{\beta}, \beta) \iff \gamma < \frac{SNR}{SNR+1}$.
- When $\gamma > 1$, $R_X(\widehat{\beta}, \beta) > R_X(\widetilde{\beta}, \beta)$ if $SNR \leq 1$.

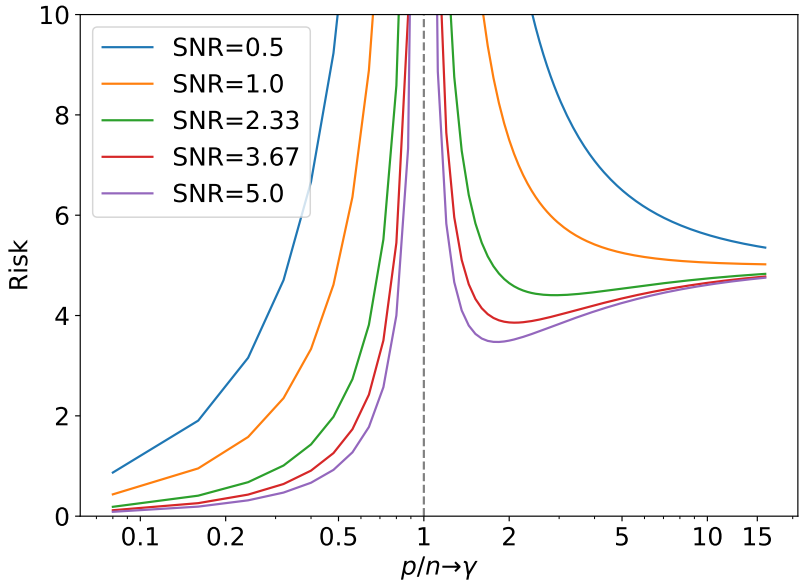
Overparametrized Asymptotics (Isotropic Features)

When $\gamma > 1, SNR > 1$, $R_X(\widehat{\beta}, \beta)$ has a local minimum at $\gamma = \frac{\sqrt{SNR}}{\sqrt{SNR}-1}$ and tends to $R_X(\widetilde{\beta}, \beta)$ from below as $\gamma \to \infty$. Recall that $SNR = \frac{r^2}{\sigma^2} = \frac{||\beta||_2^2}{\sigma^2}$ and

$$R_X(\widehat{\beta}, \beta) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2 \left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$



(a) Risk $R_X(\widehat{\beta}, \beta)$ when $\sigma^2 = 1$ and r^2 ranges from 1 to 5.



(b) Risk $R_X(\widehat{\beta}, \beta)$ when $r^2 = 5$ and σ^2 ranges from 1 to 5.

Overparametrized Asymptotics (Correlated Features)

Let $\Sigma = \sum_{i=1}^{p} s_i v_i v_i^T$ and define two probability distributions on $\mathbb{R}_{\geq 0}$:

$$\widehat{H}_n(s) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{s \geq s_i\}} \quad \text{and} \quad \widehat{G}_n(s) = \frac{1}{||\beta||_2^2}\sum_{i=1}^{p} \langle \beta, v_i \rangle^2 \mathbb{1}_{\{s \geq s_i\}}.$$

Assumption 1: $x \sim P_x$ with $x = \Sigma^{1/2} z$ and

1. $z = (z_1, ..., z_p)^T$ has independent (not necessarily identically distributed) entries with $\mathbb{E}(z_i) = 0, \mathbb{E}(z_i^2) = 1$, and $\mathbb{E}|z_i|^k \leq C_k < \infty$ for all $k \geq 2$;
2. $s_1 = ||\Sigma||_{op} \leq M$ and $\int \frac{1}{s} d\widehat{H}_n(s) < M$ for some large constant $M > 0$;
3. $\left|1 - \frac{p}{n}\right| \geq \frac{1}{M}$ and $1/M \leq p/n \leq M$.

Overparametrized Asymptotics (Correlated Features)

Under Assumption 1, we further assume that $s_p = \lambda_{\min}(\Sigma) > \frac{1}{M}$. Then, with $\gamma = p/n$, it holds with high probability that

$$R_X(\widehat{\beta}, \beta) = B_X(\widehat{\beta}, \beta) + V_X(\widehat{\beta}, \beta),$$

$$\left| B_X(\widehat{\beta}, \beta) - \mathcal{B}(\widehat{H}_n, \widehat{G}_n, \gamma) \right| \leq \frac{C\,||\beta||_2^2}{n^{1/7}},$$

$$\left| V_X(\widehat{\beta}, \beta) - \mathcal{V}(\widehat{H}_n, \gamma) \right| \leq \frac{C}{n^{1/7}},$$

where

$$\mathcal{B}(\widehat{H}_n, \widehat{G}_n, \gamma) := ||\beta||_2^2 \left[1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\widehat{H}_n(s)} \right] \int \frac{s}{(1+c_0\gamma s)^2} d\widehat{G}_n(s),$$

$$\mathcal{V}(\widehat{H}_n, \gamma) := \sigma^2 \gamma \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\widehat{H}_n(s)} \quad \text{and } 1 - \frac{1}{\gamma} = \int \frac{1}{c_0 \gamma s} d\widehat{H}_n(s).$$

Misspecified Linear Model

Now, we consider the data model

$$((x_i, w_i), \epsilon_i) \sim P_{x,w} \times P_\epsilon, i = 1, ..., n$$
$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, i = 1, ..., n,$$

where

$$\mathrm{Cov}((x_i, w_i)) = \Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw} & \Sigma_w \end{pmatrix}.$$

The out-of-sample prediction risk is defined as:

$$\begin{aligned} R_X(\widehat{\beta}; \beta, \theta) &= \mathbb{E}\left[\left(x_0^T \widehat{\beta} - \mathbb{E}\left(y_0 | x_0, w_0\right)\right)^2 | X\right] \\ &= \mathbb{E}\left[\left(x_0^T \widehat{\beta} - \mathbb{E}\left(y_0 | x_0\right)\right)^2 | X\right] + \mathbb{E}\left[\left(\mathbb{E}\left(y_0 | x_0\right) - \mathbb{E}\left(y_0 | x_0, w_0\right)\right)^2 | X\right]. \end{aligned}$$

Misspecified Linear Model (A Normal Example)

Now, we consider the data model

$$((x_i, w_i), \epsilon_i) \sim P_{x,w} \times P_\epsilon, i = 1, ..., n$$
$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, i = 1, ..., n,$$

where

$$\mathrm{Cov}((x_i, w_i)) = \Sigma = \begin{pmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw} & \Sigma_w \end{pmatrix}.$$

If (x_i, w_i) is jointly normal, then

$$\mathbb{E}\left(y_0 | x_0, w_0\right) = x_0^T \beta + w_0^T \theta$$
$$\mathbb{E}\left(y_0 | x_0\right) = x_0^T \left(\beta + \Sigma_x^{-1} \Sigma_{xw} \theta\right)$$
$$\mathbb{E}\left[\left(\mathbb{E}\left(y_0 | x_0\right) - \mathbb{E}\left(y_0 | x_0, w_0\right)\right)^2 | X\right] = \theta^T \Sigma_{w|x} \theta,$$

where $\Sigma_{w|x} = \Sigma_w - \Sigma_{xw} \Sigma_x^{-1} \Sigma_{xw}$ and

Minimum ℓ_1-Norm Regression

Question: Is it possible to consider other linear interpolators?

Minimum ℓ_1-Norm Regression

Question: Is it possible to consider other linear interpolators?

Minimum ℓ_1-norm least squares regression (Li and Wei, 2021):

$$\widehat{\beta}_{\ell_1} = \arg\min \left\{ ||b||_1 : b \text{ minimizes } ||Y - Xb||_2^2 \right\}.$$

- When $\gamma < 1$, $\widehat{\beta}_{\ell_1}$ is still the usual least squares estimator.
- When $\gamma > 1$, $\widehat{\beta}_{\ell_1}$ approaches the basis pursuit solution (Chen and Donoho, 1994):

$$\min_{b \in \mathbb{R}^p} ||b||_1$$
$$\text{subject to } Y = Xb.$$

Triple Descents in Sparse Linear Regression

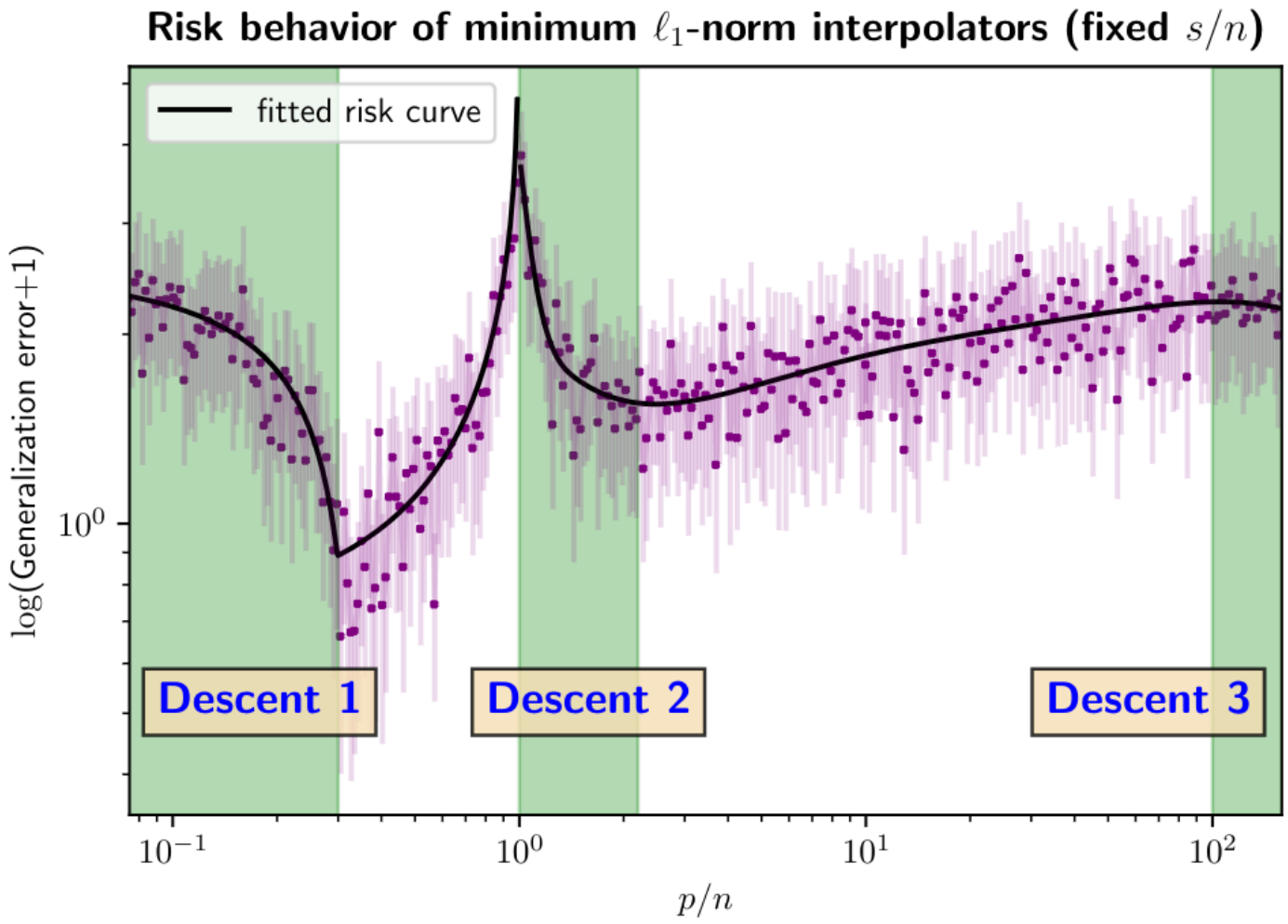


Figure 5: Triple descent in sparse linear regression (Li and Wei, 2021), where $n = 100$ is fixed, $s/n = 0.3$, and $s/n \cdot M^2 = 10$. Here, s is the sparsity level and M is the magnitude of the non-zero entries.

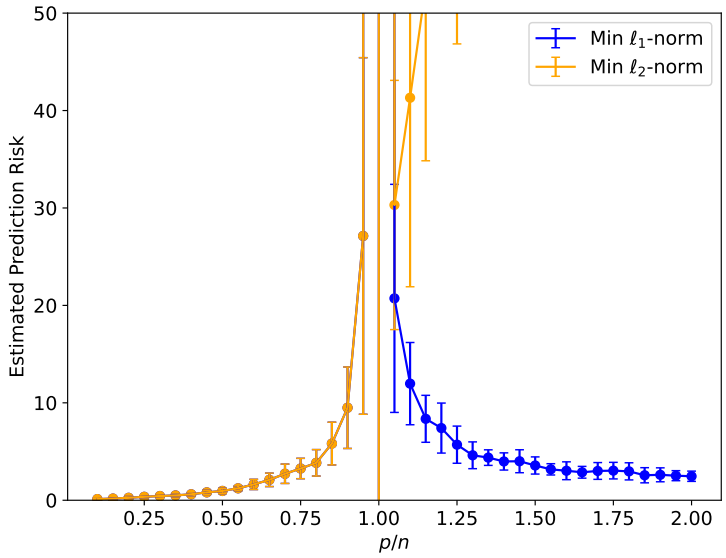
Comparisons Between Min ℓ_1 and ℓ_2-Norm Solutions

We fix $n = 100$ and generate random samples from

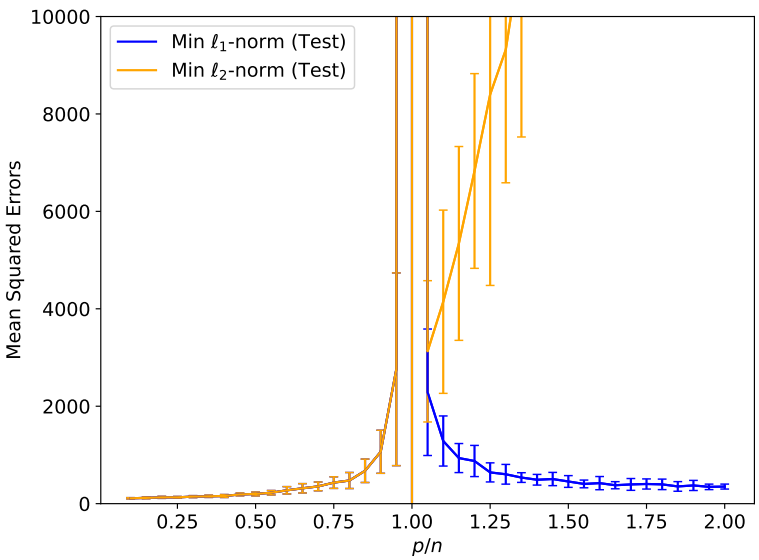
$$y = x^T\beta + \epsilon \quad \text{with} \quad (x, \epsilon) \sim \mathrm{P}_x \times \mathrm{P}_\epsilon,$$

where $P_\epsilon \sim N(0,1)$ and $P_x \sim N(0,\Sigma)$ with

$$\Sigma^{-1} = \begin{pmatrix} 1 & -0.4 & & \\ -0.4 & 1 & \ddots & \\ & \ddots & \ddots & -0.4 \\ & & -0.4 & 1 \end{pmatrix}.$$



(a) Risk $R_X(\hat{\beta}, \beta)$.



(b) Test Errors.

Further Readings

1. *Linear Regression:* "Benign Overfitting in Linear Regression" (Bartlett et al., 2020).
2. *Ridge Regression:* "The Optimal Ridge Penalty for Real-world High-dimensional Data Can Be Zero or Negative due to the Implicit Ridge Regularization" (Kobak et al., 2020).
3. *"Multiple Descent":* Li and Meng (2021).
4. *Mean-field theory:* Mei et al. (2018); Mei and Montanari (2022).
5. ...

Thank you!

Reference I

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

S. Chen and D. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

M. Geiger, S. Spigler, A. Jacot, and M. Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

D. Kobak, J. Lomond, and B. Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *The Journal of Machine Learning Research*, 21(1):6863–6878, 2020.

X. Li and X.-L. Meng. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533):353–367, 2021.

Reference II

Y. Li and Y. Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Details of "Kernel Tricks" in (1)

Recall that the ridge regression estimator $\widehat{\beta}_\lambda = \left(X^T X + n\lambda \boldsymbol{I}\right)^{-1} X^T Y$. By the Sherman–Morrison formula[1], we have that

$$
\begin{aligned}
(X^T X + n\lambda \boldsymbol{I})^{-1} X^T &= \left[\frac{1}{n\lambda}\boldsymbol{I} - \frac{1}{n^2\lambda^2} X^T \left(\boldsymbol{I} + \frac{1}{n\lambda} X X^T\right)^{-1} X\right] X^T \\
&= \frac{1}{n\lambda} X^T - \frac{1}{n^2\lambda^2} X^T \left(\boldsymbol{I} + \frac{1}{n\lambda} X X^T\right)^{-1} X X^T \\
&= \frac{1}{n\lambda} X^T - \frac{1}{n^2\lambda^2} X^T \left(\boldsymbol{I} + \frac{1}{n\lambda} X X^T\right)^{-1} \left(\frac{1}{n\lambda} X X^T + \boldsymbol{I} - \boldsymbol{I}\right) \\
&= \frac{1}{n\lambda} X^T - \frac{1}{n\lambda} X^T + \frac{1}{n\lambda} X^T \left(\boldsymbol{I} + \frac{1}{n\lambda} X X^T\right)^{-1} \\
&= X^T \left(n\lambda \boldsymbol{I} + X X^T\right)^{-1}.
\end{aligned}
$$

[1] See https://en.wikipedia.org/wiki/Sherman-Morrison_formula.